Putting Data Integration into Practice: Using Biomedical Terminologies to Add Structure to Existing Data Sources

Michael N. Cantor, M.D. and Yves A. Lussier, M.D.
Department of Biomedical Informatics, Columbia University, New York, NY 10032

A major purpose of biomedical terminologies is to provide uniform concept representation, allowing for improved methods of analysis of biomedical information. While this goal is being realized in bioinformatics, with the emergence of the Gene Ontology[a] as a standard, there is still no real standard for the representation of clinical concepts. As discoveries in biology and clinical medicine move from parallel to intersecting paths, standardized representation will become more important. A large portion of significant data, however, is mainly represented as free text, upon which conducting computer-based inferencing is nearly impossible. In order to test our hypothesis that existing biomedical terminologies, specifically the UMLS Metathesaurus® and SNOMED CT®, could be used as templates to implement semantic and logical relationships over free text data that is important both clinically and biologically, we chose to analyze OMIM™ (Online Mendelian Inheritance in Man). After finding OMIM entries' conceptual equivalents in each respective terminology, we extracted the semantic relationships that were present and evaluated a subset of them for semantic, logical, and biological legitimacy. Our study reveals the possibility of putting the knowledge present in biomedical terminologies to its intended use, with potentially clinically significant consequences.

INTRODUCTION

One of the overarching goals of the field of medical informatics is creating methods for the improved integration of biomedical data sources[1]. As the amount of biomedical data, continues to increase, integration among data sources becomes even more important. Increasingly, "it is in the correlations observed between datasets that the most interesting biological insights are found[2]." If one approaches linked biomedical sources as "networks" of information, one may also measure the value of their integration by Metcalfe's law, which states that the "value" of a network increases in proportion to the square of the number of its component nodes[3].

One of the more significant changes in biomedicine due to this information explosion involves the different perspectives being taken toward data analysis. As the post-genomic era begins, models of analysis that focus specifically on the data obtained, such as algorithms for gene clustering, are likely to be supplanted by so-called "*models-of-process*", which explain the *relationships* between genomic data and the biological pathways underlying physiologic processes[4]. Relating these processes to clinical outcomes is the next logical, though daunting, step in this process. Methods that allow for the integration of various data sources from different levels of biology may greatly facilitate this progression, perhaps potentiating the emergence of even more complex and accurate *in silico* biological models[5]. Developing these types of models is the goal of the *Molecular Medicine Matrix* (*M*[3]), currently under development at our institution. Preliminary steps in the system's development have been described previously[6].

A straightforward approach to data integration involves employing lexical methods to match terms between diverse data sources. Structured biomedical terminologies, such a SNOMED and the UMLS, may often serve as standards of measurement for determining the efficacy of a method, or as standards of linguistic knowledge as well[7]. Other authors have previously shown the feasibility of this type of data integration, using common occurrences of MeSH terms in MEDLINE references as the key to link OMIM, GENBANK, and the UMLS[8]. While this type of integration is no longer novel, we will attempt to show the feasibility of applying both these links and the semantic information they imply, in order to add a more formal internal structure to OMIM. Uncovering this internal structure may allow for even further, formal linking between OMIM and other biomedical data sources[9].

METHODS

Data sources: Currently part of the NCBI's

[a] http://www.geneontology.org

1. Process terms in "omim.txt" to obtain OMIM entry number
2. Find corresponding SNOMED CT concept ID's, UMLS CUIs through lexical matches
3. Verify semantics of matches
4. Find existing semantic relationships for source terminology concepts
5. Translate CUI/ID back to OMIM entry number
6. Find relationships among OMIM entries
7. Verify validity of subset of returned relationships

Figure 1: Schematic of methods for term and relationship extraction

system of databases, *OMIM*[b] is a comprehensive catalog of genes and genetic disorders. In addition to the free-text descriptions of each entry, it also contains information on chromosomal location, inheritance patterns, and allelic variants. The principal data source for this project was the "omim.txt" file, which contains the entire free text of the OMIM database. Each disease is represented by an OMIM code, as well as various free-text "fields", including the "Title" field that represents each disease or gene as well as its naming variants. The entire OMIM database contains over 14,000 entries, over 90% of which are autosomally inherited diseases.

UMLS. We used the 2003AA version of the UMLS Metathesaurus[c], which contains approximately 800,000 concept entries (CUIs) from over 100 biomedical vocabularies. Of note, the UMLS encompasses previous versions of both OMIM (the 1993 version), and SNOMED (version 3.5, 1998).

SNOMED CT. We used the July 2002 release of SNOMED CT[d], (SCT) which contains approximately 330,000 concepts and 1 million relationships among them. An important feature of this version of SNOMED is its incorporation of description logics, allowing for the development of inferred relationships as well[10].

Extraction of terms. A schematic of our overall methods can be seen in Figure 1. We first used Perl scripts to extract the individual entries from the "Title" and "Autosomal Variants" fields in omim.txt, resulting in approximately 70,000 individual entries. For the purpose of the following analysis, we used the extracted entries from omim.txt, the "STR" (string) field from the UMLS table MRCON, and the "Term" field from the file "sct_descriptions" as our textual sources; and OMIM numbers, UMLS CUIs, and SCT concept id's as concept identifiers. Once we had our full data set, we first attempted to match OMIM concepts exactly to the UMLS and to SCT by comparing the above textual sources. In order to expand the lexical matching, we then processed the MIM concepts and the SCT concepts with *norm*, part of the UMLS Lexical Tools[e], in order to obtain each concept's lexically normalized form. Finally, we attempted to match the normalized forms of the OMIM concepts to the normalized SCT file, as well as the MRXNS.ENG file in the UMLS. After obtaining the matches, we also verified the semantic types of both the UMLS and SNOMED entries, and removed matches between two incompatible types. We then combined the total matches to obtain the total number of OMIM concepts obtained from each source terminology.

Finding relationships. Once we had our final set of OMIM concepts, we then obtained the sets of applicable concept relations within each respective source terminology. In the UMLS, for example, we used the set of CUIs we had obtained, and found all corresponding relations present in MRREL. We repeated the same process using the SCT relationship file. Finally, we translated the CUIs and the concept id's in the relationship sets to their corresponding OMIM numbers, in order to obtain the relationships in terms of OMIM itself.

In order to perform semantic checking, exclusion lists were created for each source terminology. For the UMLS, the initial set of matching CUIs represented 98 different semantic types. Of these, 23, such as "Organism" and "Social Behavior", were put on the exclusion list. Semantic checking for SCT was more complex, as the key used for matches was the concept id, rather than the hierarchical identifier. In order to deal with this situation, we employed an ancestor-descendant table, created as part of the M^3 system, that contained the hierarchy of "is-a" links for each concept id. Through trial and error, the 3rd level down in the tree was eventually chosen as the level at which a general semantic type would be determined. At this level, there were 73 types, 28 of which, such as

[b] http://ncbi.nlm.nih.gov/omim
[c] http://umlsks.nlm.nih.gov
[d] http://www.snomed.org
[e] http://umlsks.nlm.nih.gov/

	UMLS	SNOMED CT
Concepts mapping to OMIM	7673	2794
Non-OMIM concepts related to OMIM concepts	494,660	20,493
Relationships between two OMIM concepts	44,778	352
Distinct semantic types for inter-OMIM relationships	10	8
Distinct OMIM numbers represented in relationships	2865	444

Table 1: Quantitative analysis of OMIM relationships in the source terminologies

"Vehicle" and "Non-current concept", were excluded.

After finding these sets of relations, we next attempted to prove their usefulness in terms of inference on biological concepts. As this process required manual revision of each entry, we chose a sample of 100 random relationships, drawn from both result sets, as our test set. We defined three classes of relationships: "Useful", meaning biologically plausible and not already existing in OMIM; "Questionable", meaning not proven valid based on existing evidence; and "Pre-existing", or already present in OMIM.

As an example of this entire matching process, we can take the case of OMIM entry 306900, "Hemophilia B". In the UMLS, Hemophilia B is represented by the CUI C0008533, and "MIM" is listed as one of the over 15 source vocabularies in which it is present. In SNOMED CT, Hemophilia B is represented by concept id 41788008. In the UMLS, CUI C0008533 has a relationship to 186 other CUIs, 28 of which are linked to MIM in MRSO, 62 of which have matches in our OMIM set but are not linked to MIM in MRSO, and 96 of which are not present in our OMIM set. Among these 186 relationships, 10 specific relationship types are represented. After converting the CUIs back to their corresponding MIM numbers, one example of the "SIB" (sibling) relationships we found for 306900 was with 234000, or "factor XII deficiency". In SCT, 41788008 is directly related to 11 other concept id's, none of which is in the OMIM set.

RESULTS

The main results of our analysis can be seen in Table 1. Of note, the UMLS currently only contains 240 CUIs that are directly linked to the 1993 version of OMIM. Prior to semantic checking, 7804 UMLS CUIs and 3225 SCT concepts matched out of the 70,801 different OMIM entries. Semantic checking excluded 131 UMLS CUIs and 431 SCT concepts, including 229 non-current concepts, leaving 7673 and 2794 respectively. Of these concepts, 2638 were found in both vocabularies, for a combined total of 7829 distinct OMIM codes. Of these OMIM entries, 93% were autosomal diseases. 224 of the 240 CUIs directly linked to OMIM in the UMLS were represented in the set of 7673 we obtained.

In the UMLS, we initially obtained approximately 494,000 entries in MRREL containing one or more of the 7673 CUIs we retrieved. These entries represented relationships between the CUIs in our OMIM set and approximately 64,400 CUIs that were outside the set, and inter-relationships between approximately 1,200 CUIs that were in the OMIM set. The inter-relationship set contained 10 distinct types of relationships, with the most common being the SIB relationship (approximately 80% of the total relationships). MESH was the source of approximately 50% of the CUIs in the OMIM set, while components of SCT were the sources for approximately 17%.

For the 2794 SNOMED concepts, we found 20,493 entries in the "sct_relationships" table, among 12,626 distinct concepts. Within these entries, we found 352 where both concepts were members of the OMIM set. The relationships spanned 8 different categories, with "finding site" and "associated morphology" being the most prevalent after "is-a".

Converting the concepts in the relationship sets back to their corresponding OMIM numbers allowed us to achieve our principal goal of using pre-established relationships from existing terminologies to add a semantic structure to OMIM. Since multiple OMIM numbers may be associated to each CUI or concept id, we obtained a larger set of relationships from this process. Specifically, using MRREL we obtained 44,778 entries, representing 2865 OMIM

Category	OMIM 1	Relationship	OMIM 2	Comment
Useful	**300384**-Emerin	**RB**	**188380**-Thymopoetin	Emerin mutation causes Emery-Dreifuss muscular dystrophy; Thymopoetin is active at the nicotinic Acetylcholine receptors in muscles
Questionable	**146931** – IL-9	**SIB**	**135940**-Filaggrin	IL-9 regulates lymphoid/ myeloid system. Filaggrin is an epidermal protein
Pre-existing	**187950** – Essential Thrombocythemia (ET)	**Is-a**	**6000044**-Thrombopoetin gene (THPO)	Mutation in THPO may cause ET, cited in OMIM text

Table 2: Examples of retrieved OMIM relationships

numbers, with 10 different types of semantic relationships, the vast majority of which were type "SIB". From SCT, we obtained 352 entries, representing 444 OMIM numbers, and 8 semantic types.

Though we only used a small sample for our validation step, we found 31 "useful" relationships between OMIM entries, 23 "questionable" relationships, and 46 pre-existing relationships. Sample results from this analysis can be seen in Table 2. An example of one of the "useful" relationships is a "SIB" relationship between entries 300011, ATP7A, a candidate gene involved in Menke's disease, an X-linked disorder involving cerebral degeneration, and 309550, FMR1, a gene involved in Fragile X syndrome.

DISCUSSION

The results of our study reveal the possibility of putting tools created for data integration into practice. Even with relatively basic methods, we were able to create a useful set of relationships between large biomedical terminologies and the OMIM database. More sophisticated methods for processing OMIM's full text entries probably would have permitted the extraction of several times more relationships.

One surprising result of this project was the small number of SCT relationships represented. This result may stem from several factors, including the possibility that SCT's complex semantic structure is poorly suited for analysis by simple lexical methods. The possibility of post-coordination of SCT concepts, for example, may lead to greater representation concepts at the atomic level in the terminology, rather than "complete" concepts as represented in the UMLS and OMIM. Additionally, the analysis of relationships only on a direct level may not have been appropriate to SCT's semantic structure, which may give more meaning to the entire ancestor-descendant environment of a concept. Finally, the lack of a standard, normalized file of SCT concepts such as MRXNS.ENG, as well as the inclusion of semantic information in many of the textual descriptions of SCT concepts (i.e. "Tuberculosis (disorder)"), may have led to an increased level of false negatives.

Another potential problem area for the matching methods is the approach to semantic matching. Though UMLS semantic types are relatively straightforward, there were potentially some CUIs that were falsely excluded, especially since CUIs may have more than one semantic type. The method for SCT set an arbitrary level of semantic type of a concept's parents, with an attempt to balance sensitivity and specificity, but again could have led to errors in either direction. A potential fix for this could have been to use SCT's hierarchical identifiers.

Though obtaining the set of matching concepts was an important initial first step in our study, the more significant results involve relationships among these concepts in both source terminologies. The need for even more linking in OMIM may initially seem unnecessary, as it is now part of the Entrez system at the NCBI, and its textual entries contain hyperlinks to other OMIM entries. The principal situation that our approach attempts to solve, however, looks at OMIM's linkages in a different way. For example, without text processing, extracting links to other OMIM entries requires manual searching of any specific entry, even in the Entrez system. Additionally, those links are generally straightforward, such as gene-disease, or to other diseases caused by the

same gene. Using relationships already present in the UMLS and SCT allows for the development of more complex relations, and is also a potential first step in adding additional semantic structure to the vast amount of data in OMIM. Adding semantic structure to such an impressive repository of clinical and genetic information could lead to several benefits, not the least of which would be more complex, automated queries based on the newly developed structure. Verified semantic relationships among concepts could also provide a template for further, concept-oriented coding of the extremely detailed and informative free text entries present in OMIM.

Though our methods obtained a good deal of "questionable", or perhaps yet undiscovered, relationships, more complicated methods, or even perhaps a larger or more thoroughly processed data set, could significantly increase precision. Even with improved precision, however, elucidating complex or newly discovered relationships among biological concepts will almost always require human verification. Further refinement of our methods may allow for their application to a different data source, such as a purely clinical database, for inferencing on clinical associations.

CONCLUSIONS

Though our methods are an early step in providing improved methods for information extraction from a large repository such as OMIM, what they best reveal is the possibility of using existing biomedical terminologies to find an underlying structure in biomedical information sources. More refined methods, such as those involving natural language processing or machine learning methods, should allow for improved precision and the development of a fully-verified semantic structure to a chosen set of biomedical data. With fully verified methods, especially with improved semantic and logical validation, however, these techniques could also serve to help update the UMLS to include the current version of OMIM.

We are currently exploring approaches to developing these refined methods, with the hope of applying them to different parts of the OMIM dataset. In order to provide scalability to the methods, however, the next important step will be developing automated methods for detecting "useful" relationships among concepts, with the hope of discovering linkages that may point to potentially fruitful areas of research.

Applying formal structure to a biomedical data source may lead to several advantages for researchers looking for relations between biological discoveries and disease. Like the Entrez system, which links a wide variety of sources of biomedical information, representing data and relationships in a uniform manner improves the possibility of automated inferencing and information extraction. Biological networks may not be the only important systems in the post-genomic era. Perhaps, with increasing sources of complex biomedical data, networks of well-integrated biomedical information sources may soon exemplify Metcalfe's law.

Acknowledgements: funding support provided by NLM training grant LM-07079 and NYSTAR CAT grant.

REFERENCES

1. http://www.amia.org/about/faqs.
2. Stupka E. Large-scale open bioinformatics data sources. Curr Opinion Mol Therapeutics 2002; 4:265-274.
3. Metcalfe R. Metcalfe's law: A network becomes more valuable as it reaches more users. Infoworld 1995; 17:53.
4. Voit EO. Models-of-data and models-of-process in the post-genomic era. Mathematical Biosciences 2002; 180:263-274.
5. Palsson B. In silico biology through "omics". Nature Biotechnology 2002; 20:649-50.
6. Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA. Linking biomedical language information and knowledge resources: GO and UMLS. Pac Symp Biocomput. 2003:439-50.
7. Grabar N, Zweigenbaum P. A general method for sifting linguistic knowledge from structured terminologies. Proc AMIA Symp 2000:310-4.
8. Sperzel WD, Abarbanel RM, Nelson SJ, et al. Biomedical database interconnectivity: An experiment linking MIM, GENBANK, and META-1 via MEDLINE. Proc Annu Symp Comput Appl Med Care 1991:190-193.
9. Rector A, Rogers J, Roberts A, Wroe C. Scale and context: Issues in ontologies to link health- and bio-informatics. Proc AMIA Symp 2002:642-6.
10. Spackman KA. Normal forms for description logic expressions of clinical concepts in SNOMED RT. Proc AMIA Symp 2001:627-31.